Bayesian inference of neural connectivity from simultaneous calcium imaging of a complete population of cells

(Dated: April 30, 2009)

We present Bayesian framework for inferring connectivity in a network of coupled neurons, observed simultaneously using calcium imaging.

I. MOTIVATION

The problem of reconstructing connectivity in neural circuits in the brain has recently gained much attention [EM, DTI and EEG rec prj]. In particular, amid growing evidence for the importance of collective effects in the neural networks for information processing in the brain [pop-coding olfaction & taste], the problem of understanding the pathways in which the information in biological neural networks is exchanged has swiftly become the Neuroscience's spotlight [...]. Identifying the patterns of connections between neurons in large neural circuits is a most natural way to provide empirical data about different pathways use for information processing in biological neural networks [c-elg].

In spite of constantly growing interest and increasing amount of effort directed at the problem of comprehensive neural circuit reconstruction, the methods for achieving this goal are still a subject of active development. A number of different approaches had been by now introduced [EM, pillow-retina, fMRI, DTI, B-G]. Among these, only electron microscopy may provide neural circuits reconstructions with sufficient level of resolution to catalogue connections between individual neurons in detail for large groups of neurons [Celeg]. However, even supplemented with automated data acquisition [denk] and image-processing [YM,...], electron microscopy will remain an extremely expensive approach, limited by slow imaging and vulnerable to errors in neural tracing and analysis. Reconstructions based on DTI may be performed much faster and in live subjects [DTI], however they provide very coarse resolution inadequate for mapping the fine structure of neural circuitry. Yet, recently proposed program for mapping neural connectivity statistically by utilizing large ensembles of low-resolution fluorescent probes of connectivity, obtained with stochastically expressed transsynaptic markers [B-G], may offer a fast and robust alternative for detailed anatomical reconstructions of circuitry in large neural circuits.

Alternative and complementary to anatomical reconstructions is the family of approaches inferring network connectivity from observations of neural activity [MRI, EEG, pillow-retina]. In particular, methods for inferring coarse functional connectivity maps from fMRI has been known for a while [...], and analysis of functional connectivity between ganglion cells in retina from multi-electrode array recordings has recently been presented [pillow-retina]. Although details of the relationship between functional connectivity and anatomical neural circuit structure are yet to be elaborated, empirical knowledge of functional connectivity is important both for fundamental aspects and applications. Although functional connectivity intimately entangles both circuit's anatomical structure and the type of stimuli used for its estimation (since inference of functional connectivity always relies on observation of correlations in neural activity given particular stimuli), direct links between functional and anatomical connectivity exist for special experimental conditions, as we discuss in Section X. Furthermore, direct measurement of functional and anatomical connectivity may be necessary to elucidate empirically the relationship between them. Finally, even without knowing specific mapping from functional connectivity onto anatomical circuit structure, the former still may provide sufficient predictive power for decoding or encoding of neural activity in neuro-prosthetics or neural interfaces, if the set of probed stimuli is sufficiently representative of the set of neuronal activity patterns typically encountered in applications.

Recently, great advances in the development of calcium indicators [...], delivery techniques [...], and microscopy technologies [...] have facilitated imaging of neural activity of large populations of neurons in a wide array of neural substrates. Calcium imaging provides an ultimate tool for collecting necessary data for estimating functional connectivity. Calcium imaging potentially is capable of overcoming both resolution limitation of fMRI and EEG, and cell-population size limitation of multi-electrode arrays. With calcium imaging, recordings at the level of individual cells are possible for thousands and tens of thousands of cells, while providing resolution sufficient for reconstructing individual spikes [...]. Recently, JF and LP has developed a spike-sorting tool for extracting spikes from calcium imaging data using sequential monte carlo methods. With that, spike trains may be extracted from calcium imaging with reliability sufficient for estimating functional connectivity, similar to that of direct multi-electrode recordings. That, in turns, creates an opportunity for reconstructing connectivity in neuronal microcircuits observed with calcium imaging. In this paper we develop Bayesian formalism for inferring neural connectivity in a population of neurons simultaneously observed with calcium imaging.

II. METHODS

A. Overview

Our goal for this paper will be to calculate set of connection weights $W = \{w^{kk'}\}$ between neurons in a population of N cells given simultaneous observation of activity of each of N cells with calcium imaging. We approach this problem from the positions of Bayesian inference. In this case, given set of calcium imaging observations $O = \{F^k\}$ for neurons $k = 1..N$, and posterior distribution $P(W|O) \sim P(O|W)P(W) \sim P(O, W)$, we would like to produce an estimate for $W = \{w^{kk'}\}$ given $O = \{F^k\}$. We may produce such an estimate either as the set of connection weights maximizing $\log P(W|O)$ (maximum a-posterior) or as $E[W] = \sum_W W P(W|O)$. In either case, we need to know $P(W|O)$, or $P(O|W)$ and $P(W)$, or $P(O, W)$. Of course, we do not know any of these distributions directly; however, we know how the spike trains are generated in a population of neurons, $P(H|W)$, and how spike trains result in fluorescence signal observed with calcium imaging $P(O|H, M)$. (Here, we denote with M the set of model-parameters governing calcium dynamics in neurons, and $H = \{n^k\}$ denotes the set of all spike trains in the neuronal population.) We then may consider a larger probability distribution $P(O, H, M, W)$, and approach our estimation problem as Expectation Maximization problem

$$\begin{array}{l} \text{sample } P(H|O, M, W) \sim P(O|H, M)P(H|W), \\ \max E_{P(H|O,M,W)}[\log P(H, O, M', W')] = \int dH \log P(O, H, M', W')P(H|O, M, W). \end{array} \tag{1}$$

B. Neural populations as Hidden Markov Model

Neural activity in relation to calcium imaging observations may be obviously described as a Hidden Markov Model (HMM). This HMM contains two parts - generation of spikes by individual neurons, driven by the currents injected via synaptic connections from other neurons, and the calcium dynamics in response to neuron spiking and corresponding fluorescence observations.

Specifically, we adopt canonical GLM model of coupled driven nonuniform discrete-time Poisson generators to describe spiking and interactions in a population of neurons. This model is known to capture well the statistical properties of the firing patterns of individual neurons [...].

$$\begin{array}{l} n_t^k \sim Poiss(\lambda_t^k \Delta t) \\ \lambda_t^k = g(J_t^k) = g(b_0^k + s^k \cdot X_{ext}(t) + \sum_{k'} \sum_{\tau>0} w^{kk'}(\tau) n_{t-\tau}^{k'}). \end{array} \tag{2}$$

We assume exponential rate-function $g(J) = \exp(J)$; this rate function is log-concave, thus such choice results in convex optimization problem for $w^{kk'}$ - important simplification for large volumes of neural imaging data. b_0^k and $s^k \cdot X_{ext}(t)$ are baseline and external stimulus $X_{ext}(t)$ coupling terms. Parameters $w^{kk'}$ describe coupling of activities of different neurons; we refer to these as *functional connectivity parameters*. In general, if each neuron is affected by other neurons' spikes up to K time-steps in the past, there are $\sim KN^2$ different functional connectivity weights $w^{kk'}(\tau)$. If actual neural spikes n_t^k are known, the estimation of parameters reduces to known problem such as in [...].

In the case of calcium imaging, however, we do not directly observe neural spikes (unlike in multi-electrode recordings, where neural spiking is observed directly). Instead, calcium-reporter fluorescent signal couples to neural activity via a nonlinear hidden calcium dynamics [VogPan],

$$\begin{array}{l} C_t^k \sim Normal(C_{t-1}^k + \Delta t/\tau_c^k (C_{base}^k - C_{t-1}^k) + A_n^k n_t^k, (\sigma_c^k)^2 \Delta t), \\ F_t^k \sim Poiss(S(C_t^k; \alpha_f^k, \beta_f^k, K_d^k, n_d^k)). \end{array} \tag{3}$$

Eq. (3) describes concentration of calcium in neuron k at time t, C_t^k, that would exponentially relax to the base level C_{base}^k if not for portions of calcium A_n^k injected due to spikes of neuron k. Fluorescence F_t^k, observed with calcium imaging, is represented as photon count with the mean intensity described by nonlinear transfer Hill function $S(C) = \alpha C^{n_d}/(C^{n_d} + K_d) + \beta$ [...]. Eq.(3) is governed by following parameters: calcium decay time-constant τ_c, base calcium concentration C_{base} and random fluctuations in calcium concentration σ_c, amount of calcium injected into cell upon spike A_n, scaling α and offset β for the fluorescence signal, and signal calcium saturation parameters K_d and n_d. Further, we denote $M = \{\tau_c^k, C_{base}^k, \sigma_c^k, A_n^k, \alpha_f^k, \beta_f^k, K_d^k, n_d^k\}$, and refer to these as calcium dynamics model. Note, that since calcium dynamics may be different in every cell, we assign a distinct model for each neuron, thus resulting in total of $8N$ different parameters M.

Eq.(2) and (3) define a Kth order Hidden Markov Model with state variable $X_t = (\{n_t^k\}, C = \{C_t^k\})$ and parameters M and W, with observations $O = \{F_t^k\}$. Estimation of HMM is a well developed field with a number of techniques available for efficiently sampling data, computing likelihoods, and parameters estimates [...]. In our case, however, the sheer size of the model requires development of specialized tools. In particular, for a population of N neurons we have a $2N$-dimensional state space that should be analyzed over $T \sim$ a few thousands time steps; and parameters of the model include $K \sim N^2$ coupling weights W and $8N$ calcium dynamics parameters M. For, e.g., $N \sim 100$, these numbers add up to a very formidable computational challenge.

C. Estimating properties of neural populations as Hidden Markov Model

We will be interested in inferring parameters governing neural dynamics from set of fluorescence observations O. Given that we constructed our problem as Markov Model, a natural choice for estimation framework is the Expectation-Maximization algorithm.

Expectation Maximization algorithm is an iterative estimation process, where to estimate parameters of unknown distribution $P(O, H|\theta)$, we first obtain a sample from $P(H|O, \theta_i)$, given observations and current estimate θ_i (so as to be able to compute the expected value, referred to commonly as Expectation step or E-step)

$$Q(\theta|\theta_i) = E_{P(H|O,\theta_i)} \left[\log L(O, H; \theta)\right], \tag{4}$$

and then maximize $Q(\theta|\theta_i)$ to obtain the next iterate (so called Maximization step or M-step)

$$\theta_{i+1} = \text{argmax } Q(\theta|\theta_i). \tag{5}$$

In each iteration of the algorithm, the value of the log-likelihood is guaranteed to increase and thus the algorithm is guaranteed to converge to at least a locally best model (since, obviously, log-likelihood is a quantity bounded from above). For the purposes of this paper, EM-algorithm requires implementation of two parts - efficiently sampling from HMM X_t, and efficiently maximizing HMM parameters M and W.

For HMM, sampling their state sequence is generally a very advantageous problem. For a HMM with a finite state-space, a sample state-sequence may be obtained in $O(|X|T)$ time using so called forward-backward technique, where $|X|$ is the size of the state space and T is the length of the sample sequence. This is a very advantageous scaling, given that such sampling generally implies a search over T-dimensional space of total size $|X|^T$. For continuous state-space HMM, different algorithms had been suggested relying on discretization of the continuous state-space and approximating the relationships involved in forward-backward technique [ExpProp,SMC,Fearnhead SMC,Doucet MCMC Neal's MCMCm]. In our case, we face two complications - first, our state space is hybrid, ie it is neither fully discrete nor fully continuous as one variable n_t is binary while the other C_t is continuous; and second, our state-space has extreme dimensionality $\sim 2N$. These pose substantial difficulty for all the above mentioned algorithms. In particular, reasonably-sized discretization $N_p \sim 100$ typically will be unable to sample the $2N$-dimensional state-space with reasonable accuracy, resulting in poor approximation of forward-backward relationships. In fact, sampling from HMM $H = \{X_t, t = 1..T\}$ is the most challenging part of our problem, given substantial length of recordings $T \sim$ a few thousands time points and high dimensionality of the state space X_t. For this reason, we choose a different sampling strategy relying on combination of HMM sampling techniques inside Gibbs sampling (HMM-within-Gibbs hybrid sampling algorithm).

Gibbs sampling obtains a sample from a high-dimensional distribution $P(z_1, z_2, ..., z_m)$ by repeatedly re-sampling from $P(z_i|z_1, z_2, ...z_{i-1}, z_{i+1}, ...z_m)$ either systematically or randomly. Such procedure is guaranteed to converge to the full target distribution. Given that sampling from 1D conditional distributions is easy and the chain mixes quickly, Gibbs sampling is one of the best high-dimensional sampling procedures. In our case, however, the states $(n, C)_t^k$ from the same neuron for different times t are likely to be strongly correlated, which therefore may result in slow mixing of Gibbs chain. A natural solution to this problem is to represent the full state $H = \{X_t, t = 1..T\} = \{(n, C)_t^k, k = 1..N, t = 1..T\}$ as a sequence of N block-states $x^k = \{(n, C)_t^k, t = 1..T\}$, and perform Gibbs-sampling sequentially by drawing such blocks from

$$P(x^k|H^{/k} = (x^1, ..., x^{k-1}, x^{k+1}, ...x^N), O, \theta) \tag{6}$$

using one of available HMM techniques. Then, strong correlations between "nearby" t states of the same neuron, $(n, C)_t^k$, will be handled efficiently within HMM sampler, while high-dimensionality of the complete state-space H will be handled within Gibbs sampler.

We emphasize that it is important for the purposes of this paper to obtain a sample from the joint distribution $P(H|O, \theta)$, since information about inter-neuronal couplings W is entirely encoded in fine correlations of spiking of

connected neurons over the time-scales of tens of ms - typical extent of PSP. A substantial simplification of this approach is obtained under the condition of high SNR of fluorescence observations O, in which case $P(x^k|*) \approx P(x^k|F^k)$, and the full joint distribution $P(H|O,\theta)$ factorizes $P(H|O,\theta) \approx \prod P(x^k|F^k, M^k)$. In this cases, drawing a single-neuron sample from $\{s_i^k\} \sim P(x^k|F^k, M^k)$ is sufficient, where the sample from full joint distribution may be formed simply by mix-matching spike-trains from $\{s_i^k\}$, ie $\{S_j = (s^1_{i_{1,j}}, ..., s^N_{i_{N,j}})\}$. In particular, above Gibbs sampling process in this case will converge after a single pass.

D. Sampling spike-trains from Hidden Markov Model

Sampling from HMM one-neuron at a time may be performed using any one of the technique listed above, since this problem is small enough. In case, that $w^{kk'}(\tau)$ have a particular temporal structure, ie if $w^{kk'}(\tau)$ may be described via a Markov process with respect to τ, particle filter implementation developed in [VogPan] for spike inference from calcium imaging may be immediately used. Specifically, if

$$w^{kk'}(\tau) = w^{kk'} \sum_j \exp(-\tau/\tau_j), \tag{7}$$

the coupling terms take the form

$$\sum_\tau w^{kk'}(\tau) n^{k'}(t-\tau) = w^{kk'} \sum_j h^{k'}_{j,t}, \tag{8}$$

where $h^{k'}_{j,t}$ are thus described by Markov process (with the intrinsic internal noise term $\epsilon^k_{j,t}$),

$$h^{k'}_{j,t+1} = (1 - \Delta/\tau_j) h^{k'}_{j,t} + n^{k'}_t + \epsilon^k_{j,t}. \tag{9}$$

Uniting Eq.(2),(3), (8) and (9), we arrive at an extended HMM for neural population, where relevant spike history of a spike train of neuron k is entirely encoded in its history-variable $h^k_{j,t}$. By applying particle filter from [VogPan], we may produce samples from $P(h^k_{j,t})$ for all k and t efficiently.

A more general approach to efficiently sample from our HMM is using MCMC over a set of stochastically generated grids, proposed recently by Neal, Beal and Roweis [nealMCMC]. In this approach, we obtain a sample state sequence as a limiting distribution of Markov process constructed in the following form. At every Markov step, we approximate full HMM as a discrete HMM defined on a finite grid $G = \prod G_t = \prod_{t=1..T}\{x_t^{[i]}, i = 1..m\}$. In that the sequence is selected over a finite grid of points, this method is resembling particle filter SMC. Whereas in SMC the grid is constructed via evolving a swarm of particles, here the points at each t are drawn independently from a prior distributions $\{\rho_t(x_t)\}$. The sequence of states is thus selected using forward-backward procedure over this grid, according to probability

$$P(x_{1:T}|y_{1:T}) \sim P(x_0) \prod_{t=1}^{t=n-1} P(x_t|x_{t-1}) \prod_{t=0}^{t=n-1} \frac{P(y_t|x_t)}{\rho_t(x_t)}. \tag{10}$$

This sequence is subsequently incorporated into the stochastic grid for the next Markov step as the first point $x_t^{[1]}$ into G_t for each t. [Neal] show that the limiting distribution of this Markov Chain, indeed, is exactly

$$P(x_0) \prod_{t=1}^{t=n-1} P(x_t|x_{t-1}) \prod_{t=0}^{t=n-1} P(y_t|x_t). \tag{11}$$

Given this latter property of [nealMCMC] procedure, this sampler is exact, whereas SMC approach is discretization-biased [Doucet, Neal]. Specifically, because sequence of states in SMC is selected over a particular realization of grid G, as constructed in the forward pass of SMC, the final samples are distributed with probability

$$P(\{x_t\} = P(x_0) \prod_{t=1}^{t=n-1} P(x_t|x_{t-1}) \prod_{t=0}^{t=n-1} P(y_t|x_t)/Z(G), \tag{12}$$

where $Z(G) = \sum_{\{x_t\}\in G} P(x_0) \prod_{t=1}^{t=n-1} P(x_t|x_{t-1}) \prod_{t=0}^{t=n-1} P(y_t|x_t)$. Given errors in specific estimation of the normalization constant $Z(G)$, for a particular forward-pass selected grid of points G, SMC sample is thus biased. By performing

selection of state-sequences over a "chain" of grids G^i, drawn randomly from some well-controlled distribution ρ^T, the problem of biasing the sample via wrong estimation of $Z(G)$ is fully circumvented.

For $\rho_t(x_t)$, it is possible to choose potentially any distribution with sufficiently large support; however, in order to achieve faster convergence of Markov chain, we use the proposal distribution $\rho_t(x_t) \sim P_{PF}(x_t)$ from the particle filter such as in [VogPan]. In more details, we construct $\rho_t(x_t) = \rho_t(n_t, C_t) = \rho_t(n_t)\rho_t(C_t)$, where $\rho_t(C_t)$ is a mixture of Gaussians with variance $\sigma \approx var\left[C_t^{[i]} - C_t^{[j]}\right]$, centered on particles from the particle filter $C_t^{[i]}$, and $\rho_t(n_t)$ is a Bernoulli distribution with certain spiking probability. It is also possible to choose $rho_t(x_t) \sim P(F_t|C_t)\rho_t(n_t)$, without previously estimating $P(x_t|*)$ via particle filter [VogPan].

With [nealMCMC] approach, we may draw spike-trains directly (instead of history-variable representation $h_{j,t}^k$) and without making specific assumptions about the temporal structure of $w^{kk'}(\tau)$. After we have obtained a sample of states $H = \{x^k\}$ via HMM-within-Gibbs sampling techniques, we can evaluate the expected value of the log-likelihood and, subsequently, perform its maximization aiming to find new set of parameters, better describing our observations.

E. Estimating parameters of the Hidden Markov Model

Having sampled from $P(H|O, M, W)$, we can evaluate the expected value of the HMM log-likelihood given different set of parameters (M', W'), i.e. perform the E-step. Within the framework of EM, we now need to find the best new set of parameters, or perform the M-step. The optimization problem we need to solve has the following structure:

$$E[\log P] \sim \sum_k E_H[\log P_{[Ca]}(F^k|x^k, M_k)] + \sum_k E_H[\log P_{GLM}(x^k|H^{/k}, W)], \tag{13}$$

where $P_{[Ca]}(F^k|x^k, M_k; W, H)$ is the likelihood of obtaining the sequence of observations F^k given kth neuron calcium dynamics model M_k, and spike trains $H = \{x^k\}$ and the connectivity W (see [VogPan] for details), and

$$\begin{array}{l} P_{GLM}(x^k|H^{/k}, W) = \sum_t \left(n_t^k \log g(J_t^k|W, H) - (1 - n_t^k) g(J_t^k|W, H)\Delta t\right), \\ J_t^k = b_0^k + s^k \cdot X_{ext}(t) + \sum\limits_{k'} \sum\limits_{\tau>0} w^{kk'}(\tau) n_{t-\tau}^{k'}, \end{array} \tag{14}$$

is the standard GLM log-likelihood for a sequence of spikes given input currents (2).

For large number of neurons N, we are facing optimization problem that is very large - we need to maximize expected log-likelihood with respect to $8N$ parameters in M and N^2 parameters in W. Fortunately, this optimization problem may be solved efficiently. In particular, estimation of $M = \{M_k\}$ parameters may be performed separately for each neuron, since calcium dynamics are not interrelated and, given x^k, are also uncoupled from W. Thus, optimization with respect to M only involves solving N 8-dimensional problems, which may be done quickly (see [VogPan] for details). Optimization with respect to W is in N^2 variables; however, by construction, GLM log-likelihood $P_{GLM}(x^k, H^{/k}, W)$ is convex in W and so we can use standard efficient algorithm to maximize in W, e.g. such as gradient ascent or Newton method.

F. Specific implementation

In our specific implementation of the above process, we break the problem of inference of connectivity in a population of neurons in three steps [DIAGRAM] - we first estimate for each neuron k the model of its calcium dynamics M_k given the fluorescent observations F^k and the estimate of injected current $J_t^k = \sum_{k'} w^{kk'}(\tau) n_{t-\tau}^{k'}$ (given current estimate of $w^{kk'}(\tau)$ and $n_t^{k'}$; at first iteration, $w^{kk'} \equiv 0$). Estimation of the models M_k is performed on a subset of calcium imaging data containing $\sim 10 - 100$ spikes, by first drawing a sample spike-train and associated calcium concentration from $P(n_t^k, C_t^k|F_{1:T}^k, J_t^k, M_k)$, and then maximizing the expected value of the log-likelihood with respect to M_k in an iterative EM-loop. Obtaining the samples may be done using nealsMCM or SMC algorithm in [VogPan]. The latter is somewhat faster since estimation of M_k only requires $P(n_t^k, C_t^k|F_{1:T}^k, J_t^k, M)$, and these may be acquired faster with SMC (note that the spike-history terms $h_{j,t}^k$ may be calculated during this step as well).

It is advantageous to perform estimation of M_k in a separate EM-subloop given that this problem is decoupled from estimation of W, it may be performed using smaller subset of calcium imaging data, and that quite satisfactory models M_k may be obtained without updating W or spike-trains of other neurons $x^{k'}$ (per our observations). Thus, performing M_k estimation in a EM-subloop allows one to arrive at substantially better spike-train samples H even before the first estimation of W is done. Estimation of W requires acquisition of very long samples x^k, necessary to accurately solve large joint-optimization in W, which constitutes the bulk computational cost of this method. Thus,

prior estimation of M allows to avoid recomputing long samples x^k for estimation of W while M_k-estimates are converging.

In the second step, given calcium dynamics models M_k for each neuron, we obtain a full spike-train sample for each k from $P(x^k|H^{/k}, M, W)$, one neuron at a time, in a HMM-within-Gibbs-sampler loop using neal's MCMC. Note that in high SNR regime, given our notes above, obtaining HMM-sample for each neuron may be performed independently from the others and, typically, is only required once (since $P(x^k|H^{/k}, M, W)$ are essentially defined by $P(x^k|H^{/k}, M, W) \approx P(x^k|F^k)$). It is also possible to use SMC to perform selection of x^k (assuming discretization bias may be tolerated), or completely circumvent this step by using $h^k_{j,t}$ calculated in step one (if we assume known Markovian temporal form of $w^{kk'}(\tau)$, as discussed in the previous section).

Finally, given joint sample of $H = \{x^k\}$, we perform estimation of connectivity matrix W by solving standard max-log-likelihood problem for corresponding GLM [Pan*]. The above steps one through three may be then repeated until convergence is observed in the connectivity matrix W and spike-train samples H.

Note, that in our specific implementation we obtain a sample of x^k only using spike-trains of other neurons to calculate and estimate the injected currents J^k_t. This process corresponds to only accounting for information about past neural activity in the population when determining whether neuron k should spike at time t - $n^k_t \sim P(n^k_t|\{n^{k'}_{t'}\}_{t'=1..t})$. In principle, taking into account the impact of neuron k spiking at time t on other neurons via GLM couplings, along with the information available about activity of other neurons in the future, allows for a better sample from $P(n^k_t|\{n^{k'}_{t'}\}_{t'=1..T})$ [Pillow]. While we recognize importance of this note, specific implementation of this process may be cumbersome, and in this work we do not implement such improved procedure for generating samples x^k, although we do plan to incorporate it in the above framework in the future.

G. Simulation inference of neural connectivity for neural population

To test the performance for inference of neural connectivity in a neural population, we simulated such inference in conditions close to such expected in real data. Specifically, we solve connectivity inference for a network of simulated spiking neurons, constructed closely following empirical data known about the real neural networks in the cortex. We assume a population of neurons that are spontaneously firing action potentials. While it is known, that the functional connectivity weights in general do not properly reflect anatomical connectivity in a circuit, we will show that for the above system of spontaneously firing neurons the direct correspondence, indeed, exists, and anatomical connectivity is properly recovered via functional connectivity weights.

Functional connectivity may fail to faithfully represent anatomical circuit structure if false correlations are present between different neurons, induced e.g. by common inputs, or if the dynamics of neural population is entirely concentrated on a low-dimensional subspace of the full configurational space H. Note that these two statements are, in a sense, different ways of stating the same condition: if activity of different neurons is tightly correlated, their dynamics is concentrated on a low-dimensional plane; and vice-versa - concentration of dynamics onto a low-dimensional plane will be perceived as correlation in activity of different neurons. (In turn, low dimensionality of the neural dynamics may be caused by different factors, including common input, small subset of command neurons driving the circuit, or even emergent property of a network.) Low dimensionality of the neural dynamics results in that the inference problem Eq.(13) becomes underdetermined, i.e. there may exist directions in W along which connectivity is not constrained by activity data (i.e. directions orthogonal to the subspace of all observed neural activity configurations), or is poorly constrained. This, naturally, leads to W being poorly defined along these directions. The necessary condition for good correspondence between functional connectivity weights W and anatomical connectivity, therefore, is *full-dimensionality* of the observed neural dynamics. In case of spontaneously firing system of neurons this condition is, in fact, satisfied by many independent neuron-ignitions, thus, fully sampling possible directions in the configurational space H. If spontaneously active preparation by itself fails to display sufficient degree of independence between randomly firing neurons (e.g. if low-dimensionality of the activity subspace is the emergent property of studied circuit), such pattern may be induced by randomly activating subsets of neurons via ChR2 [...] or glutamate uncaging.

We also note that the correlations induced by secondary and so on synaptic transmissions (such as when neuron A results in firing of neuron B, which in turn results in firing by neuron C), are all properly resolved in GLM-fitting process via the so called explaining-away process. In other words, because we do not just identify correlations between neural firings with the functional connectivity weights $w^{kk'}$, but instead statistically fit a model of neural interactions, if found weights between neurons A and B, and B and C are sufficient to explain the correlation between A and C, the weight connecting A and C will not appear in the model - the correlation between A and C was "explained away" by correlations between A and B, and B and C. By this, the multi-synaptic firing patterns do not confuse our estimation process.

We prepared a small network of $N = 50$ neurons randomly sparsely connected, and firing stochastically with the base firing rate of about 5Hz. Each neuron was modeled with GLM as a linear-nonlinear driven Poisson spike generator, as described above

$$\begin{array}{l} P(n^k_{t+1}|\{h_{t+1}\}) = \text{Poiss}(\lambda^k_{t+1}\Delta t) \\ \lambda^k_{t+1} = g(J^k_t) = g(b_0 + \sum_{\tau>0} w^{k,k'}_\tau n^{k'}_{t-\tau}), \end{array} \tag{15}$$

with exponential transfer function $g(J) = \exp(J)$. We assumed no external modulating input $X_{ext}(t)$.

The network was divided into excitatory and inhibitory components. Neurons in such components were either entirely excitatory or inhibitory; i.e., all connections outgoing from such neuron were either all simultaneously positive (excitatory) or negative (inhibitory). Excitatory neurons were randomly connected with each other and the inhibitory neurons with probability $f_c = 0.1$ (observed probability for a local connection between nearby neurons in the cortex [**]). The synaptic weight of each connection v, as defined by max EPSP amplitude, were generated from exponential distribution with mean $0.5\mu V$ [**] (we neglected here the "heavy tail" of the distribution of synaptic weights observed in some datasets [**]). While synaptic weights are typically measured in PSP units of μV, in GLM model connections are measured in log-rate units of Eq.(15). In other words, GLM weights describe the *change in probability of neuron k to fire given neuron k′ has firing before*. By utilizing this definition, we may convert a PSP weight v_{EPSP} for a neuron V_{base} below spike-triggering threshold into the GLM weight as $v_{GLM} \approx v_{EPSP}/V_{base}$. In other words, V_{base}/v_{EPSP} spikes are required to push neuron over the threshold. Given the definition of the firing rate in (15), this leads to the following equation for the log-rate couplings $w^{kk'}$

$$w^{kk'} = \ln(-\ln(\exp(-f^k\Delta t) - v^{kk'}_{EPSP}/V_{base})/\Delta t/f^k), \tag{16}$$

where f^k is the base "stochastic" firing rate of neuron k.

20% of all neurons were taken to be inhibitory [**]. Interneurons were randomly connected among themselves and to excitatory neurons with the same frequency $f_c = 0.1$ as above. The strength of the inhibitory connection was drawn from the exponential distribution with mean chosen to balance excitatory and inhibitory currents, and to achieve the final firing rate close to the base firing rate $f = g(b_0)$. Connection strengths were thus converted to log-rate units using Eq.(16), where $w^{kk'}$ was finally multiplied by -1 to reflect inhibition.

To generate a sequence of spikes in this population, we simulated activity of the network forward in time computing currents injected into each cell from all previous spikes of all neurons. Each spike was assumed to inject the same PSP waveform, described by a temporal filter h_{PSP} [DIAGRAM] modeled as the difference of two exponentials with the rise time of $1ms$ and decay time of $10ms$ [**]

$$J^k_{t,inject} = \sum_{k'\neq k} w^{kk'} \sum_{\tau>0} h_{PSP}(\tau) n^{k'}_{t-\tau}. \tag{17}$$

(Given $1ms$ time step of our simulation, and the fact that a signal in a local cortical circuit of the size of $\sim 1mm$ would suffer $\leq 1ms$ time lag, we neglect delays in this simulation.) Additionally, each neuron exhibited refractory current with waveform h_{REFR} [DIAGRAM] modeled with exponential with decay time of $5ms$ [**]

$$J^k_{t,refr} = \Omega^k \sum_{\tau>0} h_{REFR}(\tau) n^k_{t-\tau}. \tag{18}$$

The network was then thus simulated forward in time with step of $\Delta t = 1ms$. All neurons were assumed to be electro-physiologically similar (i.e. have the same PSP profiles, Ω and base firing rate, etc.) Spikes were generated at each time according to Bernoulli distribution $n^k_t = \text{Bernoulli}[g(J^k_{t,spont} + J^k_{t,refr} + J^k_{t,inject})\Delta t]$.

Given a sequence of spikes, the fluorescence observations were generated using the setup in [VogPan]. Parameters for the model were chosen in accordance to our experience of analyzing a few actual cells [**]. Specifically, calcium decay time constant was $\approx 0.25s$, the ratio of per-spike calcium influx to stochastic fluctuations in calcium concentration was $\approx 3:1$, and the background of calcium concentration was chosen to be about 30% of per-spike calcium influx, thus corresponding to typical relatively low SNR of 1:3. Photon count per measurement was taken to be $\approx 3 \cdot 10^4$. The population of cells was generated with these parameters while assuming that all parameters may vary by at least 30% from cell to cell. Thus, we generated fluorescence for each cell using its unique cell of parameters, and produced sampled observations at frame-rate of 33Hz or 66Hz. [PARAMS-TABLE]

We simulated from $T = 300s$ to $T = 600s$ of observation data. To determine the necessary population observation time, we used Fisher information. For simplicity, consider case where $K = 1$ and a perfect knowledge of spike-histories (i.e. such that was not corrupted by inference errors from calcium imaging data). Then, we may write Fisher

information for GLM as

$$C^{-1} = \frac{\partial(-\ln P)}{\partial w^{kl}\partial w^{k'l'}} = -\delta_{kk'}\sum_t \left(n_t^k n_{t-1}^l n_{t-1}^{l'} \left(-\frac{g'(J_t^k)^2}{g(J_t^k)^2} + \frac{g''(J_t^k)}{g(J_t^k)} \right) - \Delta t(1-n_t^k) n_{t-1}^l n_{t-1}^{l'} g''(J_t^k) \right). \tag{19}$$

Assuming exponential transfer function $g(J)$ and weak coupling between spikes, this may be rewritten as

$$\begin{aligned} C^{-1} &= \delta_{kk'}(T\Delta t) P(n_t^k = 0, n_{t-1}^l = 1, n_{t-1}^{l'} = 1) E[g(J_t^k) | n_t^k = 0, n_{t-1}^l = 1, n_{t-1}^{l'} = 1] \\ &\approx \delta_{kk'}(T\Delta t)(f^l \Delta T)(f^{l'} \Delta T) f^k \sim (\Delta T f)^2 (T\Delta t f) M. \end{aligned} \tag{20}$$

Here $T\Delta t$ is the observation time, f is the typical firing rate, and ΔT is "the coincidence time", i.e. the interval of time within which the spike from neuron A affects the spike probability of neuron B (typically, the extent of PSP). M is some $O(1)$ residual matrix, which we assume may be viewed as as a matrix made up of N diagonal random blocks of size $N \times N$. For successful determination of the weights, the variance C should be smaller than the typical scale of the weights W. Then, using [**] for the typical value of the smaller singular value of a random matrix of size $N \times N$, we obtain

$$T\Delta t \sim \sqrt{N}/(\langle W \rangle f)(\langle \Delta T \rangle f)^2. \tag{21}$$

E.g., for $N \approx 100$, $\langle W \rangle \approx 1$, $f \approx 5Hz$ and $\Delta T \approx 10ms$, we obtain $T \approx 800s$.